Journal of the Korea Academia-Industrial cooperation Society
Vol. 24, No. 2 pp. 11-20, 2023

https://doi.org/10.5762/KAIS.2023.24.2.11
ISSN 1975-4701 / eISSN 2288-4688

준지도 학습 모델을 사용한 차량 내부 네트워크에서의 이상 징후 탐지

이현로, 홍성우, 이승열, 하재철*
호서대학교 컴퓨터공학부

Anomaly Detection in Vehicle Network Using Semi-Supervised Learning Model

Hyunro Lee, Seongwoo Hong, Seungyeol Lee, Jaecheol Ha*
Division of Computer Engineering, Hoseo University

요 약 현재 자동차 내부의 전자 제어 장치(Electronic Control Unit, ECU)간의 통신을 위해 CAN(Controller Area Network) 프로토콜을 많이 사용하고 있다. 하지만 CAN 프로토콜은 메시지 암호화 및 발신자 인증과 같은 보안 기능을 가지고 있지 않아 인가되지 않은 데이터 주입이나 서비스 거부 공격(Denial of Service, DoS) 등과 같은 사이버 보안 위협에 취약하다. 따라서 최근에는 자동차의 CAN 네트워크를 보호하기 위한 인공 지능 기반의 침입 탐지 시스템(Intrusion Detection System, IDS)에 대한 연구가 활발하게 진행되고 있다. 본 논문에서는 먼저 CAN 버스의 데이터 트래픽에 대한 메시지 주입 공격을 탐지할 수 있는 지도 학습(supervised learning)을 사용하는 딥러닝 모델인 DCNN(Deep Convolutional Neural network) 기반의 이상 탐지 모델을 구현하였다. 또한, 지도 학습 모델은 학습용 데이터 셋이 많아야 한다는 한계점을 지적하고 이를 보완하기 위해 준지도 학습(semi-supervised learning)을 사용한 딥러닝 모델인 GAN(Generative Adversarial Network) 기반의 이상 탐지 모델을 제안한다. 제안하는 준지도 학습 기반의 이상 탐지 모델은 기존 지도 학습 모델에서 약 20만 개의 데이터로 학습하던 것을 1,000개의 데이터만으로도 서비스 거부 공격과 스푸핑 공격을 99%이상 탐지할 수 있어 효율적인 차량용 이상 징후 탐지 시스템으로 사용할 수 있다.

Abstract The CAN (Controller Area Network) protocol is used widely for communication between ECUs (Electronic Control Units) in a vehicle network. On the other hand, the CAN protocol is vulnerable to cyber security threats, such as unauthorized data injection and DoS (Denial of Service) attacks, because it does not have security functions, such as message encryption and sender authentication. Therefore, research on an artificial intelligence-based IDS (Intrusion Detection System) for protecting the CAN network has been actively conducted. This paper reports an anomaly detection model based on a DCNN (Deep Convolutional Neural Network), a deep learning model using supervised learning that can detect message injection attacks on data traffic on CAN buses. The supervised learning model requires a large number of training data sets. This paper proposes an anomaly detection model based on GAN (Generative Adversarial Network), a deep learning model using semi-supervised learning, to compensate for this advantage. The proposed anomaly detection model based on semi-supervised learning can be used as an efficient vehicle anomaly detection system because it can detect more than 99% of denial-of-service and spoofing attacks with only 1,000 data instead of learning with about 200,000 data in the existing supervised learning model.

Keywords : Cloud Storage, Client-Side Deduplication, Identification Attack, Data Deletion Process, Cloud Security

본 논문은 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다.
(No. 2021RIS-004)
*Corresponding Author : Jaecheol Ha(Hoseo Univ.)
email: jcha@hoseo.edu
Received January 5, 2023 Revised January 30, 2023
Accepted February 3, 2023 Published February 28, 2023

1. 서론

최근에는 자율 주행차, 커넥티드카와 같이 다양한 정보 기술을 차량에 접목하여 사용자에게 편의성을 제공하는 방향으로 발전하고 있다. 하지만 다양한 IT 기술들이 접목되면서 외부 공격자가 CAN(Controller Area Network), LIN(Local Interconnected Network), FlexRay, MOST(Media Oriented Systems Transport) 와 같은 차량 내부 네트워크에 접근할 수 있는 경로가 증가하여 차량의 사이버 보안 위협 문제가 대두되고 있다.

차량 내부 네트워크 통신에서 많이 사용되고 있는 CAN은 데이터가 폐쇄된 네트워크에서만 유지된다는 가정하에 설계되었기 때문에 네트워크를 보호하기 위한 메시지 암호화나 발신자 인증과 같은 보안 메커니즘이 적용되어 있지 않다[1]. 따라서 외부 공격자는 CAN 버스의 데이터 트래픽을 도청할 수 있으며 인가되지 않은 데이터를 CAN 버스에 주입하여 차량 시스템을 제어할 수 있다. 예를 들어 공격자는 엔진 속도, 기어, 브레이크 등과 같은 특정 기능과 관련된 CAN ID로 메시지를 주입하여 엔진의 속도를 올리거나 기어를 변경할 수 있다. 이와 같은 오류로 차량이 오동작을 일으키면 매우 치명적인 인명 피해가 발생할 수 있으며, CAN 버스에서 도청한 메시지를 통해 차량의 내부 정보와 개인 정보도 탈취할 수 있다.

CAN 버스에 대한 보안 문제가 대두되면서 다양한 CAN 보안 메커니즘들이 제안되었다. 방화벽을 통해 외부 인터페이스에서 자동차용 전자 장치를 분리하거나 엄격한 접근 통제를 적용해 신뢰할 수 있는 부분만 자동차 내 시스템에 접근하도록 허용하는 통제 보안 메커니즘이 제안되기도 하였다. 또한, 발신자 인증, 데이터 무결성, 최신 암호화를 적용하여 자동차 내부 네트워크를 보호하는 암호 보안 메커니즘에 대한 연구도 진행되고 있다[2,3]. 최근에는 차량 내·외부에서 발생하는 데이터를 분석하고 머신러닝 기반의 이상 탐지 기법을 사용하여 보안 위협을 검출하는 연구가 활발히 진행 중이다.

구체적인 연구 사례를 보면, A. Taylor 등은 LSTM(Long Short-Term Memory) 네트워크 기반의 모델을 사용하여 실제 차량에서 수집된 CAN 트래픽 로그에서 각 송신자로부터 시작되는 다음 데이터 단어를 예측하여 이상 신호를 탐지하는 이상 탐지 시스템을 제안하였다[4]. C. Wang 등은 계층적 시간 메모리(HTM, Hierarchical Temporal Memory)를 이용한 분산 이상 탐지 시스템을 제안하였다. 여기에서는 이상 탐지 점수를 계산하기 위해 HTM 알고리즘과 로그 손실 함수를 기반으로 예측 변수를 설계하였다[5].

또한, H. Song 등은 확장된 CAN 메시지에서 29비트의 CAN ID와 CAN 트래픽의 시퀀스를 2차원 형태의 공간적 로컬 상관관계로 재구성하고 지도 학습 기반의 합성곱 신경망(CNN) 딥러닝 모델을 사용하여 이상 탐지 시스템을 제안하였다[6]. 이 외에도 다양한 방법론으로 CAN 버스 시스템을 위한 이상 탐지 시스템에 대한 연구가 활발히 진행 중이다[7].

본 논문에서는 메시지 주입 공격이나 서비스 거부 공격(Denial of Service, DoS)과 같은 사이버 공격으로부터 CAN 버스를 보호하기 위해 효율적인 딥러닝 기반의 침입 탐지 시스템을 제안한다. 이를 위해 먼저 지도 학습(supervised learning)에 기반한 딥러닝 모델인 DCNN(Deep Convolutional Neural Network) 기반의 이상 탐지 모델을 구현하였다. 그러나 지도 학습 모델은 높은 이상 탐지 성능을 보이지만 약 20만 개 이상의 라벨링된 데이터가 사용되었다. 특히, 차량 내부의 네트워킹 능력에 비해 너무 크고 복잡한 이상 징후 탐지 모델은 학습과 추론에서 많은 시간이 걸린다는 점도 고려하여야 한다. 따라서 자동차의 차종에 따른 이상 징후 학습 데이터가 충분하지 않은 경우에도 사용할 수 있는 효율적인 딥러닝 모델이 필요하다.

본 논문에서는 이러한 지도 학습 모델의 한계점을 보완하기 위해 준지도 학습(semi-supervised learning)을 사용한 딥러닝 모델인 GAN(Generative Adversarial Network) 기반의 이상 탐지 모델을 제안한다. 제안하는 준지도 학습 기반의 이상 탐지 모델은 1,000개 정도의 라벨링된 데이터만으로도 DoS 공격이나 Spoofing 공격을 99%이상 탐지할 수 있어 효율적인 차량용 이상 징후 탐지 시스템으로 사용할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서 CAN 프로토콜의 개념과 CAN 프로토콜에서 취약한 메시지 주입 공격에 대해 설명한다. 3장에서는 학습 방법에 따른 이상 탐지 모델을 제안하고 4장에서는 차량용 이상 탐지 실험을 수행하여 성능을 평가하고 모델별 장·단점을 비교 분석한다.

2. 배경 지식

2.1 CAN 프로토콜

자동차 내부 네트워크인 CAN은 1985년 Bosch사에

서 처음 개발되었으며, 차량 내에서 호스트 컴퓨터 없이 마이크로컨트롤러나 전자 장치들이 서로 통신하기 위해 설계되었다. 1993년에 ISO 국제 표준 규격(ISO 11898)으로 제정되었으며 1994년부터 CANopen, Device Net 등 상위 레벨 프로토콜로 표준화 되었다.

CAN은 식별자의 길이에 따라 두 가지 버전으로 구분된다. 표준 버전인 CAN 2.0A는 11비트 식별자를 기반으로 노드를 구별하고, 확장 버전인 CAN 2.0B는 29비트 식별자를 기반으로 노드를 구별한다. CAN 2.0A 컨트롤러는 표준 CAN 포맷의 메시지만 송·수신이 가능하며, 확장 CAN 포맷 메시지를 수신하면 식별자를 인식할 수 없어 해당 메시지를 수용하지 못하고 거부한다. 반면에 CAN 2.0B 컨트롤러는 표준 CAN 포맷의 메시지 ID를 인식할 수 있으므로 모든 포맷에 대해 모두 송·수신이 가능하다[8].

CAN 메시지 포맷은 데이터 프레임(data frame), 리모트 프레임(remote frame), 에러 프레임(error frame), 오버로드 프레임(overload frame) 4개의 프레임 타입으로 정의하고 있다. 리모트 프레임은 수신할 노드에서 원하는 메시지를 전송할 수 있는 송신 노드에게 전송을 요청할 때 사용된다. 에러 프레임은 메시지의 에러가 감지되었을 때 시스템에 알릴 목적으로 사용된다. 오버로드 프레임은 메시지의 동기화를 목적으로 사용한다.

CAN 통신에서 중요한 데이터 송·수신은 데이터 프레임을 사용하여 이루어진다. 데이터 프레임 구조는 Fig. 1과 같으며 CAN 데이터 프레임에서 각 필드의 의미는 다음과 같다.

- SOF(Start of Frame): 한 개의 dominant 비트로 구성되어 있으며 메시지의 처음을 지시하고 모든 노드의 동기화를 위해 사용된다.
- Arbitration Field : 11비트 또는 29비트의 크기를 갖는 ID와 1비트의 RTR(Remote Transmisson Request) 비트로 구성된다.
- Control Field : 2비트의 IDE(IDentifier Extension) 비트, 4비트의 데이터 길이 코드(DLC, Data Length Code)로 구성되어 있으며 R0은 Reserved 비트(Extended CAN 2.0B R0, R1)이다.
- Data Field : 최대 8바이트까지 사용 가능하며 데이터를 저장하는 데 사용된다.
- CRC(Cyclic Redundancy Check) : SOF에서부터 데이터 필드까지의 비트열을 이용해 생성한 15비트의 CRC 시퀀스와 하나의 열성 비트('r')의 CRC 델리미터로 구성되어 있으며 메시지 상의 에러 유무를 검사하는 데 사용된다.
- ACK(Acknowledgment): 1비트의 ACK 슬롯과 하나의 ACK 델리미터('d')로 구성되어 있으며 임의의 노드에서 올바른 메시지를 수신하게 되면 ACK 필드를 받는 순간 ACK 슬롯의 값을 우성 비트('d')로 설정해 버스 상에서 계속 전송하게 된다.
- EOF (End of Frame) : 7개의 'r' 비트로 구성되어 있으며 메시지 끝을 알리는 목적으로 사용된다.

2.2 메시지 주입 공격

CAN 프로토콜에서 발생하는 보안 문제는 송신자 인

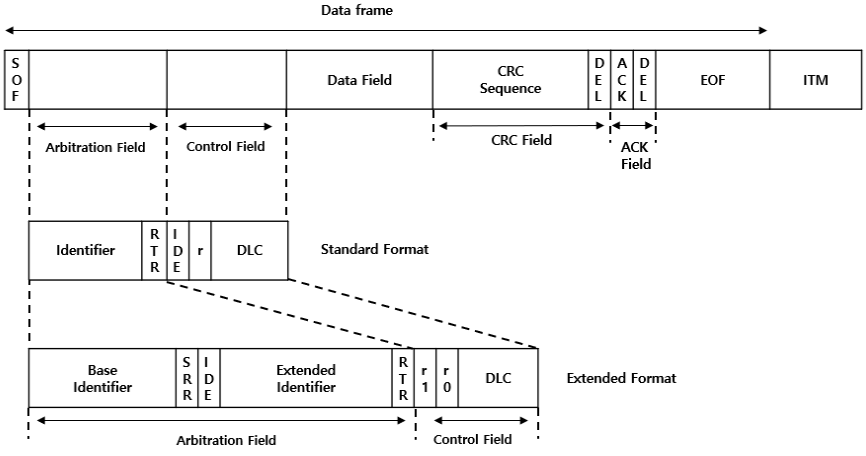

Fig. 1. Structure of CAN data frame

증 및 암호화와 같은 보안 기능이 없기 때문이다. 기본적으로 CAN은 브로드캐스트 기반 버스 네트워크이고 인증이 없기 때문에 어떤 노드든 버스에 연결할 수 있으며 모든 메시지를 수신할 수 있다. 따라서 공격자는 CAN 버스 데이터를 쉽게 스니핑할 수 있다. 즉, 내부 데이터가 쉽게 노출됨으로써 공격자가 대상 차량의 CAN 데이터를 분석하고, 이후 메시지 주입 공격을 시도하여 대상 차량을 제어할 수 있게 된다.

구체적으로 S. Checkoway 등은 공격자가 CAN 버스에 인가되지 않은 메시지를 주입하기 위해 접근할 수 있는 다양한 공격 지점을 제안하기도 하였다[9]. 이들은 차량의 공격 지점을 크게 물리적 접근과 무선 접근으로 분류했다. 물리적 접근에는 공격자가 OBD-II 포트를 통해 CAN 버스에 간접적인 물리적 액세스를 하거나 자동차의 MP3, 오디오와 같은 엔터테인먼트 시스템에 악성 코드를 인코딩한 노래 파일 또는 CD로 접근할 수 있었다.

무선 접근에는 블루투스, RFID(Radio-Frequency Identification), WiFi, 3G 및 LTE와 같은 장·단거리 무선 채널을 통해 원격으로 접근할 수 있었다. 이러한 공격 채널들은 대상 차량의 CAN 버스에 대한 진입점으로 악용될 수 있다. 다음 Fig. 2는 일반적인 물리적, 무선 접근을 통한 메시지 주입 공격 시나리오를 도시한 것이다.

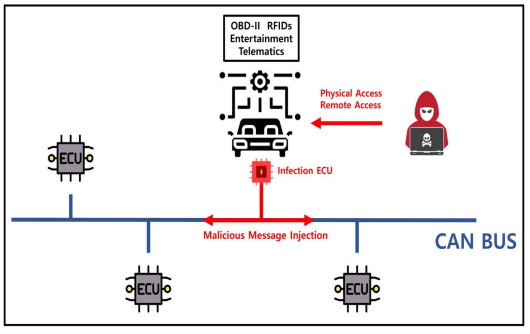

Fig. 2. Message injection scenarios into CAN bus

상기한 방법으로 CAN 버스의 노드 중 하나를 장악하면 인가되지 않은 메시지를 주입하여 차량을 제어할 수 있다. 메시지 주입 공격 종류는 재전송 공격, 스푸핑(spoofing) 공격, 서비스 거부 공격, 퍼지(fuzzy) 공격 등이 있다. 각 공격 방식을 설명하면 다음과 같다.

- 재전송 공격 : 프로토콜 상에서 유효한 메시지를 복사한 후 재전송함으로써 정당한 사용자로 위장할 수 있는 공격 기법으로 CAN 프로토콜은 브로드캐스트 방식을 사용하기 때문에 CAN 버스의 장악한 노드를 통해 모든 메시지를 수집하고 재전송하여 그대로 동작시킬 수 있다.
- 스푸핑 공격 : 승인받은 사용자인 것처럼 시스템에 접근하거나 네트워크상에서 허가된 노드로 가장하여 접근제어를 우회하는 공격 기법으로 수집된 메시지를 이용하여 변조된 메시지를 주입하여 오동작을 일으킨다.
- 서비스 거부 공격 : 시스템을 악의적으로 공격해 해당 시스템 자원을 부족하게 하여 원래 의도된 용도로 사용하지 못하게 하는 공격으로 CAN 버스 상에서 가장 높은 우선순위를 가진 ID를 지속적으로 주입하여 자원을 사용하지 못하게 한다.
- 퍼지 공격 : 소프트웨어 취약점을 발견하기 위해 프로그램이나 단말 혹은 시스템에 비정상적인 입력 데이터를 보내는 방법으로 CAN ID와 Data Field를 고려하지 않고 랜덤한 값을 이용하여 자동차의 취약점을 유발시킨다.

3. 딥러닝 기반 이상 탐지 모델

3.1 지도 학습 이상탐지

CAN Bus 상에서 메시지 주입 공격에 대한 보안성을 강화하기 위해 CAN 버스의 데이터 트래픽의 시퀀스를 공간적인 로컬 상관관계로 재구성한 뒤 ResNet34 (Residual Network34) 모델을 사용한 지도 학습 기반의 이상 탐지 모델을 제안한다.

ResNet은 모델의 계층이 깊어 질수록 기울기가 소실되는 문제 때문에 학습이 잘 이루어지지 않는 현상을 극복하기 위해 처음 고안되었다[10]. ResNet은 Shortcut connection을 이용한 Residual learning 기법을 통해 계층이 깊어짐에 따른 기울기 소실 문제를 해결하였다. 기존 신경망의 학습 목적은 입력(input) 값 x를 목표 값 y로 매핑(mapping)하는 함수 $H(\,)$를 찾는 것이다. 따라서 $H(x)-y$를 최소화하는 방향으로 학습을 진행한다. 하지만 Residual learning 기법은 네트워크의 출력 값이 x에 근접하도록 잔차 값$F(x)=H(x)-x$를 최소화하는 방향으로 학습을 진행한다. 다음 Fig. 3은 ResNet에서 사용한 Residual learning 기법을 도시한 것이다.

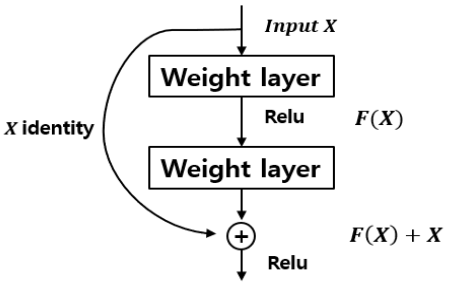

Fig. 3. ResNet shortcut connection

이때 네트워크의 출력(output)이 x가 되도록 매핑하는 것이 아닌 Fig. 3과 같이 마지막에 x를 더해 네트워크의 출력은 0이 되도록 매핑하여 최종 출력이 x가 되도록 학습한다. 따라서 각 계층이 깊어져도 최소 기울기는 1 이상의 값을 가지기 때문에 기울기 소실 문제를 해결할 수 있다.

ResNet34는 Shortcut connection으로 만든 Identity block과 Convolution block으로 구성되어 있다. Identity block은 네트워크의 출력 $F(x)$에 x를 그대로 더하는 것이고, Convolution block은 다운 샘플링(down sampling)을 하기 위해 x를 1x1 Convolution 연산을 거친 후 $F(x)$에 더한다. Fig. 4는 Identity block과 Convolution block의 구조를 보여주며 ResNet34의 전체 구조는 34개의 layer를 Identity block과 Convolution block으로 구성되어 있다.

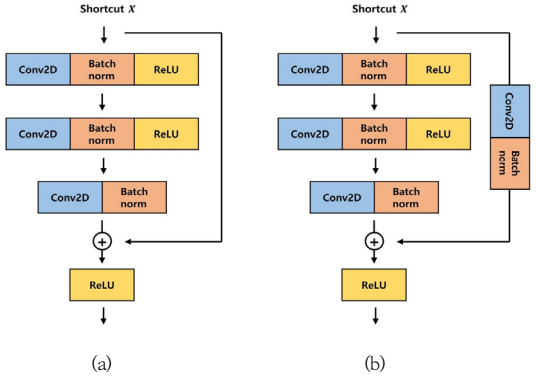

Fig. 4. (a) Identity block and (b) Convolution Block

상기한 모델로 지도 학습을 하기 위해 CAN 버스의 데이터 트래픽에서 CAN ID, DLC Field의 Hex 값을 추출한다. 각 Hex 값은 Fig. 5와 같이 15 비트(CAN ID 11bit + DLC 4bit)로 표현한 후 시간적 관계를 고려하여 15개의 트래픽을 묶어 15x15의 2차원 데이터로 변환하였다.

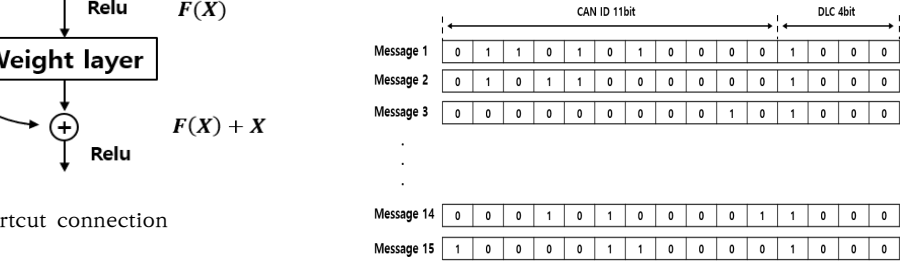

Fig. 5. Reconstructed 2D array using CAN traffic data

변환된 데이터는 시간적인 특징을 담고 있기 때문에 Convolution 과정에서 엣지(edge) 부분의 손실을 최소화하기 위해 제로 패딩(zero padding)을 통해 Fig. 6과 같이 24x24의 이미지로 만들어지게 된다.

CAN 트래픽 이미지는 15개의 데이터 시퀀스 중에서 'R' 플래그만 포함하고 있으면 Normal로 라벨링(labeling)하고 1개 이상의 'T' 플래그를 포함하고 있으면 해당 이미지는 Abnormal로 라벨링하여 학습을 진행하였다. 본 논문에서는 ResNet34와 동일한 구조로 학습을 진행하였으며 해당 모델로 테스트한 성능을 기존의 연구 결과와 비교하였다.

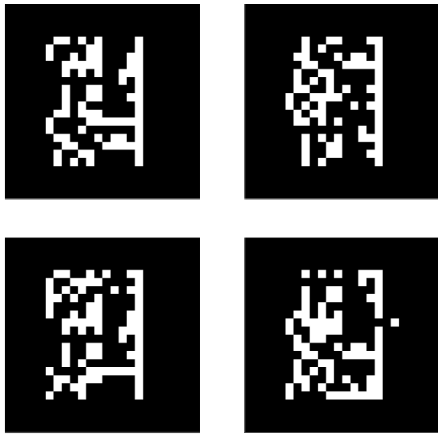

Fig. 6. Reconstructed 2D images with zero padding

3.2 준지도 학습 기반 이상탐지

지도 학습 기반의 이상탐지 시스템은 학습 데이터의 특징과 패턴으로 학습하는 방법으로 많은 라벨링된 데이

터(labeled data)가 필요하다. 하지만 실제 환경에서는 많은 라벨링된 데이터를 수집하는 것이 어려우며 잘못된 라벨링된 데이터로 학습할 경우 학습이 제대로 되지 않는 문제가 발생할 수 있다.

이러한 문제점을 해결하기 위해 준지도 학습 기반의 이상 탐지 시스템도 많은 연구가 진행되고 있다[11,12]. 준지도 학습은 라벨링된 데이터를 이용한 지도 학습과 라벨링되지 않은 데이터(unlabeled data)를 이용한 비지도 학습을 결합하여 학습 능력을 개선한다. 소수의 라벨링된 데이터는 다수의 라벨링되지 않은 데이터를 학습 과정에서 가이드 역할을 해주고 다수의 라벨링되지 않은 데이터로부터 라벨링된 데이터는 자연스럽게 일반화가 가능해진다.

본 논문에서는 준지도 학습 기반의 이상탐지 시스템을 제안하기 위해 SGAN(Semi-Supervised GAN) 모델을 사용하였으며 SGAN의 구조는 Fig. 7과 같다. 기존 GAN은 생성자와 판별자를 통하여 서로를 보완하는 방향으로 학습하여 결과적으로 판별자가 분류할 수 없는 생성자를 얻는 것이 큰 목적이지만 SGAN은 생성자와 판별자를 통해 더 좋은 판별자를 얻는 것이 목적인 모델이다[13].

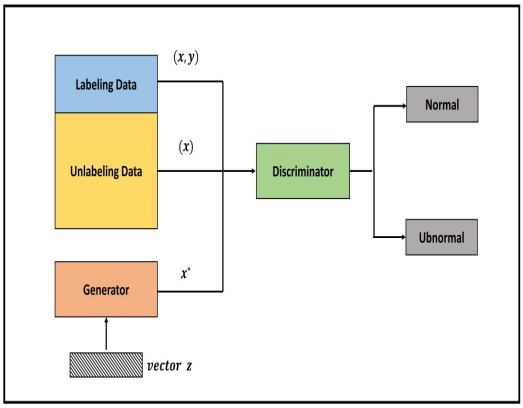

Fig. 7. SGAN architecture

SGAN을 이용한 이상 탐지를 위해 CAN 트래픽 데이터의 전처리는 지도 학습 기반의 이상 탐지와 동일하게 적용하였다. SGAN의 Discriminator에 사용되는 라벨링된 데이터에는 Normal, Abnormal 데이터모두 사용하였고 라벨링되지 않은 데이터에는 Normal 데이터만을 사용하였다. 또한 SGAN 기존 모델에서 Supervised Discriminator와 Unsupervised Discriminator의 Convolution 계층을 추가하였고 Learning rate scheduler를 적용하여 모델 학습에서 빠르게 최적화되게 하였다. 본 논문에서 학습에 사용된 Supervised Discriminator, Unsupervised Discriminator, Generator의 구조는 각각 Fig. 8과 같으며 학습 알고리즘은 다음과 같다.

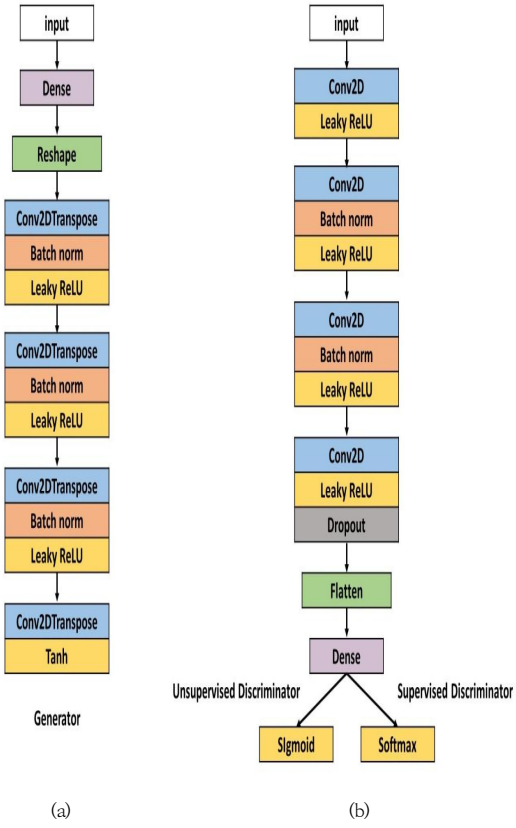

Fig. 8. (a) SGAN Generator and (b) SGAN Discriminator

① Supervised Discriminator를 학습

라벨링된 데이터에서 진짜 샘플 (x,y)의 랜덤 미니배치를 얻고 $D((x,y))$를 계산한 후 다중 분류 손실을 역전파하여 가중치 $\theta^{(D)}$를 업데이트하고 손실을 최소화한다.

② Unsupervised Discriminator를 학습

라벨링되지 않은 데이터에서 진짜 샘플 x의 랜덤 미니배치를 얻고 $D(x)$를 계산한 후 이진 분류 손실을 역전파하여 가중치 $\theta^{(D)}$를 업데이트하고 손실을 최소화한다. 이후 랜덤한 벡터 z의 미니배치로 가짜 샘플 $G(z)=x^*$의 미니 배치를 생성하고 $D(x^*)$을 계산한

후 이진 분류 손실을 역전파하여 가중치 $\theta^{(D)}$를 업데이트하고 손실을 최소화한다.

③ Generator를 학습

랜덤한 벡터 z의 미니배치로 가짜 샘플 $G(z)=x^*$의 미니 배치를 생성하고 $D(x^*)$을 계산한 후 이진 분류 손실을 역전파하여 가중치 $\theta^{(D)}$를 업데이트하고 손실을 최소화한다.

본 논문에서는 지도 학습 기반의 이상 탐지 모델과 마찬가지로 조정된 SGAN 모델로 테스트를 진행하였고 테스트한 성능을 분석 비교하였다.

4. 실험 및 이상 탐지 모델 성능 평가

상기한 지도 학습, 준지도 학습 기반의 이상 탐지 실험 및 성능 평가를 위해 이전 연구[6]에서 사용한 데이터 셋을 사용하였다. 해당 데이터 셋은 메시지 주입 공격이 수행되는 동안 실제 차량에서 OBD-II 포트를 통해 추출된 데이터 셋으로 Normal 데이터, DoS 공격 데이터, Fuzzy 공격 데이터, Spoofing(Gear, RPM) 공격 데이터를 포함하고 있다[14]. 각 공격 데이터 셋은 Table 1과 같은 방법으로 추출되었으며 데이터는 Timestamp, CAN ID, DLC, DATA, Flag Field로 구성되어 있다.

Table 1. Description of attack type

Attack Type	Description
DoS	Inject CAN ID '0000' every 0.3ms
Fuzzy	Inject random CAN ID DATA every 0.5ms
Spoofing Gear	Injects a message of a specific CAN ID related to gear information every 1ms
Spoofing RPM	Injects a message of a specific CAN ID related to RPM information every 1ms

본 논문에서 사용한 학습용 데이터의 구성을 나타낸 것이 Table 2이다. 먼저 정상 패킷은 약 98만 개를 사용하였으며, 공격 형태에 따라 정상 패킷과 공격용 패킷이 섞인 약 360만개에서 440만개 정도의 패킷을 사용하였다. 각 패킷은 15개씩 묶어 15x15의 이진 이미지를 구성한 뒤 제로 패딩을 거쳐 24x24의 이미지로 확장하여 사용하였다.

Table 2. Learning data according to attack type

Attack Type	Packets	ResNet34 (Images)		SGAN (Images)	
		N	A	U(N)	L(N+A)
DoS	3,665,771	203,691	57,748	65,932	10/100/1000
Fuzzy	3,838,860	202,423	68,248	65,932	10/100/1000
Spoofing Gear	4,443,142	202,341	100,558	65,932	10/100/1000
Spoofing RPM	4,621,702	202,223	110,199	65,932	10/100/1000
N : Normal A : Abnormal U : Unlabeled L : Labeled					

지도 학습 기반의 이상 탐지 모델은 정상 데이터와 각 공격 데이터의 80%를 합쳐 학습에 사용하였고 20%는 모델 성능 평가에 사용되었다. 예를 들어 DoS 공격에 대한 지도 학습 모델에는 약 180,000장의 라벨링된 Normal 이미지와 54,000장의 라벨링된 Abnormal 이미지를 사용하여 학습하였다. ResNet34 모델에서 사용한 하이퍼 파라미터는 256 배치 크기(batch size), 5 에포크(epochs), Adam 옵티마이저(optimizer), 0.001 학습률(learning rate)을 사용하였다.

준지도 학습 기반의 이상 탐지 모델 SGAN에 사용된 Discriminator는 학습 데이터로 라벨링되지 않은 데이터, 라벨링된 데이터, $G(z)$를 사용한다. 라벨링되지 않은 데이터는 정상 데이터만을 사용하고 라벨링된 데이터는 80%의 공격 데이터에서 10개, 100개, 1000개를 사용하였다. 라벨링된 이미지에서 Normal 이미지 Abnormal 이미지의 비율은 50:50으로 설정하였다. 또한, 모델의 성능 평가를 위해서는 학습에 사용되지 않은 공격 데이터의 20%를 사용하였다.

SGAN 모델의 하이퍼 파라미터는 256 배치 크기, 60 에포크, Adam 옵티마이저를 사용하였고 학습 최적화를 위해 Step learning rate scheduler를 사용하였다. Step learning rate scheduler는 일정 step마다 학습률에 gamma를 곱해주는 방식으로 학습 초반에는 큰 학습률인 0.01로 설정하여 빠르게 적응시키고 최적값에 가까워질수록 학습률을 줄여 학습에 도움을 주었다.

지도 학습 기반의 모델과 준지도 학습 기반 모델의 성능 평가를 위해 평가지표로 Recall, Precision, F1-score, AUC(Area Under the Curve)를 사용하였다.

Precision은 실제로 Abnormal 데이터를 모델이 Abnormal로 인식한 데이터의 수이며 수식은 다음과 같다.

$$\frac{TruePositives}{TruePositives + FalsePositives} \quad (1)$$

Recall은 모델이 Abnormal로 예측한 데이터 중 실제로 Abnormal인 데이터의 수이며 수식은 다음과 같다.

$$\frac{TruePositives}{TruePositives + FalseNegatives} \quad (2)$$

F1-score는 Precision과 Recall의 조화 평균으로 데이터 셋의 클래스 분포가 고르지 않을 때 모델의 분류 성능을 측정하는 데 사용되며 수식은 다음과 같다.

$$2 \times \frac{\mathrm{P}recision \times Recall}{\mathrm{P}recision + Recall} \quad (3)$$

AUC는 ROC 곡선 면적에 기반한 값이다. ROC 곡선은 False Positive Rate가 변할 때 True Positive Rate가 어떻게 변하는지를 나타내는 곡선으로 AUC score는 ROC 곡선 아래의 면적을 나타낸다. AUC score는 1에 가까울수록 분류 모델의 분류 성능이 우수하다고 판단한다.

다음 Table 3은 이전 연구[6]에서 수행한 Inception ResNet 모델의 구조를 변경하여 테스트한 성능을 보여준다. 이 모델에서는 확장 CAN을 사용하여 29비트의 CAN ID를 전처리한 뒤 모델의 입력으로 사용하였다. 또한, 이 연구에서 사용한 평가 지표는 Recall, Precision, F1-score으로서 모두 99% 이상의 높은 탐지율을 보였다.

다음 Table 4는 지도 학습 기반의 모델의 추가 연구를 위해 본 실험에서는 CAN ID 11비트와 DLC 4비트를 같은 방법으로 전처리하고 제로 패딩을 적용한 뒤 ResNet34 모델을 사용하여 학습에 사용하지 않은 테스트 데이터로 테스트한 평가지표이다.

Table 3. Performance of Reduced Inception-ResNet[6]

	Precision	Recall	F1-score	AUC
DoS	1.0000	0.9989	0.9995	-
Fuzzy	0.9995	0.9965	0.9980	-
Gear	0.9999	0.9989	0.9994	-
RPM	0.9999	0.9994	0.9996	-

Table 4. Performance of our ResNet34

	Precision	Recall	F1-score	AUC
DoS	1.0000	0.9997	0.9998	0.9998
Fuzzy	0.9989	0.9764	0.9875	0.9879
Gear	0.9996	0.9950	0.9973	0.9973
RPM	0.9990	0.9989	0.9990	0.9991

본 논문에서 구현한 ResNet34 지도 학습에 의한 이상 징후 탐지에서도 이전 연구 [6]의 결과와 거의 유사한 성능을 나타내었다. 특히, 지도 학습 기반의 이상 탐지 모델들은 모든 공격에서 98% 이상의 AUC-score와 F1-score를 도출하였다.

그러나 지도 학습 모델은 높은 이상 탐지 성능을 보이지만 약 20만 개 이상의 라벨링된 데이터가 필요하고 사용한 DCNN 모델의 구조가 차량 내부 네트워크에 맞추기에는 너무 크고 복잡하여 학습과 추론에서 많은 시간이 걸린다는 단점이 있다.

따라서 본 논문에서는 라벨링된 학습 데이터의 양이 충분하지 않은 상태에서 이상 징후를 탐지하기 위한 방법으로 준지도 학습법인 SGAN을 구현하였다. 다음 Table 5, Table 6, Table 7은 준지도 학습 기반의 SGAN 모델을 각각 라벨링된 데이터 10장, 100장, 1000장을 사용하여 학습한 모델을 테스트한 평가 지표이다.

표에서 보는 바와 같이 준지도 학습 기반의 이상 탐지 모델에서 10장이나 100장 정도의 라벨링된 데이터 이미지로는 좋은 탐지 성능을 볼 수 없었다. 그러나 약 1,000장 정도의 라벨링된 이미지를 사용하면 이상 징후 탐지 시스템으로 사용할 수 있을 정도의 성능을 나타내었다. 특히, DoS 공격이나 Spoofing 공격에 대해서는 99% 이상의 탐지 성능을 나타내고 있다. 다만, Fuzzy 공격에 대해서는 약 90%정도의 탐지 성능을 보여 이에 대한 원인 분석 및 데이터 셋 추가 문제는 고려해야 할 것이다. 그럼에도 불구하고 제안하는 준지도 학습 모델은 기본적으로 구조가 가볍고 단순하여 차량 내부 네트워크의 이상 탐지 모델로 적합한 것으로 여겨진다.

Table 5. Performance of our SGAN using 10 images

	Precision	Recall	F1-score	AUC
DoS	0.7114	0.2043	0.3174	0.5860
Fuzzy	0.3536	0.2414	0.2869	0.5349
Gear	0.2815	1.0000	0.4393	0.5036
RPM	0.5641	0.2200	0.3165	0.5769

Table 6. Performance of our SGAN using 100 images

	Precision	Recall	F1-score	AUC
DoS	1.0000	0.9743	0.9870	0.9871
Fuzzy	0.9071	0.2371	0.3760	0.6138
Gear	0.6948	0.9886	0.8160	0.9098
RPM	0.7196	0.7514	0.8581	0.8757

Table 7. Performance of our SGAN using 1000 images

	Precision	Recall	F1-score	AUC
DoS	1.0000	0.9857	0.9928	0.9929
Fuzzy	0.9289	0.8771	0.9023	0.9255
Gear	0.9986	0.9886	0.9935	0.9940
RPM	0.9817	0.9943	0.9879	0.9935

5. 결론

최근에는 자동차와 관련한 다양한 IT 기술이 접목되고 차량에 접근할 수 있는 채널이 늘어남에 따라 차량 내부 네트워크 CAN에 대한 사이버 보안 위협 이 증대되고 있다. 특히 CAN 버스의 노드를 장악하고 인가되지 않은 메시지를 주입하는 메시지 주입 공격은 차량을 오작동시켜 큰 인명 피해가 발생할 수 있다. 따라서 본 논문에서는 인가되지 않은 메시지를 사전에 탐지할 수 있는 딥러닝 기반의 이상 탐지 모델을 설계하였다. 제안하는 지도 학습 기반의 이상 탐지 모델은 높은 탐지율을 보여주지만 라벨링된 데이터 셋이 충분하지 못할 경우 적용이 어렵다는 단점이 있다.

따라서 지도 학습 모델은 갖는 근본적인 한계점을 보완할 수 있는 준지도 학습 기반의 이상 탐지 모델을 제시하였다. 제안하는 준지도 학습 모델은 약 1,000개 정도의 적은 양의 라벨링된 데이터로도 99%이상의 탐지율로 DoS 공격이나 Spoofing 공격 징후를 탐지할 수 있음을 실험으로 증명하였다. 향후에는 준지도 학습 모델이 Fuzzy 공격에 대한 탐지율이 다소 낮게 나타나는 원인을 분석하고 이에 대한 대응 방안에 대한 연구가 필요하다.

References

[1] O. Avatefipour and H. Malik, "State-of-the-art survey on in-vehicle network communication (CAN-Bus) security and vulnerabilities," IJCSN Journal Vol. 6, Issue 6, 2018.
DOI: https://doi.org/10.48550/arXiv.1802.01725

[2] A. Herrewege, D. Singelee, and I. Verbauwhed, "CANAuth - A Simple, Backward Compatible broadcast authentication protocol for CAN bus," In ECRYPT workshop on Lightweight Cryptography. p. 20, 2011.

[3] B. Palaniswamy, S. Camtepe, E. Foo, and J. Pieprzyk, "An efficient authentication scheme for intra-vehicular controller area network," IEEE Transactions on Information Forensics and Security 15, pp. 3107-3122, 2020.
DOI: https://doi.org/10.1109/TIFS.2020.2983285

[4] A. Taylor, S. Leblanc, and N. Japkowicz, "Anomaly detection in automobile control network data with long short-term memory networks," In 2016 IEEE International Conference on Data Science and Advanced Analytics(DSAA), IEEE, pp. 130-139, 2016.
DOI: https://doi.org/10.1109/dsaa.2016.20

[5] C. Wang, Z. Zhao, L. Gong, L. Zhu, Z. Liu, and X. Cheng, "A distributed anomaly detection system for in-vehicle network using HTM," IEEE Access 6, pp. 9091-9098, 2018.
DOI: https://doi.org/10.1109/access.2018.2799210

[6] H. Song M. Woo, and H. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," Vehicular Communications, Vol. 21, Issue C, p. 100198, 2020.
DOI: https://doi.org/10.1016/j.vehcom.2019.100198

[7] S. Lokman, A. Othman, and M. Abu-Bakar, "Intrusion detection system for automotive Controller Area Network(CAN) bus system: a review," EURASIP Journal on wireless Communications and Networking. pp. 1-17, 2019.
DOI: https://doi.org/10.1186/s13638-019-1484-3

[8] N. Lakhal, O. Nasri, L. Adouane, J. Slama, "Controller area network reliability: overview of design challenges and safety related perspectives of future transportation systems," IET Intelligent Transport Systems, Vol. 14 Issue 13, pp. 1727-1739, 2020.
DOI: https://doi.org/10.1049/iet-its.2019.0565

[9] S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, S. Savage, and T. Kohno, "Comprehensive experimental analyses of automotive attack surfaces," Proceedings of the 20th USENIX conference on Security(SEC'11), p. 7792, 2011.
DOI: https://dl.acm.org/doi/10.5555/2028067.2028073

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
DOI: https://doi.org/10.1109/cvpr.2016.90

[11] C. Chen, Y. Gong, and Y. Tian, "Semi-supervised learning methods for network intrusion detection," In 2008 IEEE international conference on systems, man and cybernetics, IEEE, pp. 2603-2608, 2008.
DOI: https://doi.org/10.1109/icsmc.2008.4811688

[12] Y. Dong, K. Chen, Y. Peng, and Z. Ma, "Comparative Study on Supervised versus Semi-supervised Machine Learning for Anomaly Detection of In-vehicle CAN Network," In 2022 IEEE 25th International Conference on Intelligent Transportation Systems(ITSC), pp. 2914-1919, 2022.
DOI: https://doi.org/10.48550/arXiv.2207.10286

[13] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation,"

IEEE 15th International Symposium on Biomedical Imaging(ISBI'18), pp. 1038-1042, 2018.
DOI: https://doi.org/10.1109/ISBI.2018.8363749

[14] H. Song and H. Kim, "Car-Hacking Dataset for the intrusion detection," Available at http://ocslab.hksecurity.net/Datasets/car-hacking-dataset (Accessed 30 December 2018).

이 현 로(Hyunro Lee) [준회원]

• 2017년 3월 ~ 현재 : 호서대학교 컴퓨터공학부 학부과정

〈관심분야〉
자동차 보안, 부채널 공격, 양자내성 암호, 머신러닝

홍 성 우(Seongwoo Hong) [준회원]

• 2017년 3월 ~ 현재 : 호서대학교 컴퓨터공학부 학부과정

〈관심분야〉
부채널 공격, 암호학, 정보보호, 인공지능 보안

이 승 열(Seungyeol Lee) [준회원]

• 2018년 3월 ~ 현재 : 호서대학교 컴퓨터공학부 학부과정

〈관심분야〉
인공지능 보안, 부채널 공격, 양자 내성 암호

하 재 철(Jaecheol Ha) [종신회원]

• 1989년 2월 : 경북대학교 전자공학과 (학사)
• 1993년 8월 : 경북대학교 전자공학과 (석사)
• 1998년 2월 : 경북대학교 전자공학과 (박사)
• 1998년 3월 ~ 2007년 2월 : 나사렛대학교 정보통신학과 교수
• 2007년 3월 ~ 현재 : 호서대학교 컴퓨터공학부 교수
• 2009년 1월 ~ 현재 : 한국산학기술학회 이사
• 2013년 1월 ~ 현재 : 한국정보보호학회 수석부회장
• 2023년 1월 ~ 현재 : 국제차세대융합기술학회 부회장

〈관심분야〉
암호학, 네트워크 보안, 부채널 공격, 머신러닝